lncDIFF: a novel distribution-free method for differential expression analysis of long non-coding RNA

Qian Li[1], Xiaoqing Yu[2], Ritu Chaudhary[3], Robbert JC Slebos[3], Christine H. Chung[3], Xuefeng Wang[2]*

[1]Health Informatics Institute, University of South Florida, Tampa, FL 33612, USA

[2]Department of Biostatistics and Bioinformatics, Moffitt Cancer Center, Tampa, FL 33612, USA

[3]Department of Head and Neck-Endocrine Oncology, Moffitt Cancer Center, Tampa, FL 33612, USA.

* To whom correspondence should be addressed. Tel: +1 813-745-6710; Fax +1 913-745-6107; Email: xuefeng.wang@moffitt.org

ABSTRACT

Motivation

Long non-coding RNA expression data has been increasingly used in finding diagnostic and prognostic biomarkers in cancer studies. Existing differential analysis tools for RNA sequencing does not effectively accommodate low abundant genes, as commonly observed in lncRNA. We propose a novel and robust statistical method lncDIFF to detect differential expressed (DE) genes without assuming the true density on normalized counts.

Results

lncDIFF adopts the generalized linear model with zero-inflated exponential quasi likelihood to estimate group effect on normalized counts, and employs the likelihood ratio test to detect differential expressed genes. The proposed method and tool is suitable for data processed with standard RNA-Seq preprocessing and normalization pipelines. Simulation results illustrate that lncDIFF detects DE genes with more power and lower false discovery rate regardless of the data pattern. The analysis on a head and neck squamous cell carcinomas study also confirms that lncDIFF has better sensitivity in identifying novel lncRNA genes with relatively large fold change and prognostic value.

Availability and Implementation

lncDIFF is an R package available at https://github.com/qianli10000/lncDIFF.

Supplementary Information

Supplementary Data are available at Bioinformatics online.

INTRODUCTION

Long noncoding RNAs (lncRNAs) are transcripts longer than 200 nucleotides with no or limited protein-coding capability. It is estimated that, in the human genome, there are at least four times more lncRNA genes than protein-coding genes [1]. Currently, there are more than 14,000 human lncRNAs annotated in GENCODE (https://www.gencodegenes.org/). Overall, lncRNA genes have fewer exons, lower abundance and are under selective constraints compared to protein-coding genes. LncRNAs are involved in diverse regulatory mechanisms and in some critical pathways. For example, they can act as scaffolds to create higher-order protein complexes, as decoys to bind sequester transcription factors, and as guides of protein-DNA interactions [2-4]. Emerging evidence suggests that lncRNA serve as essential regulators in cancer cell migration and invasion, as well as in other cancerous phenotypes [5, 6]. Therefore, lncRNAs are becoming attractive potential therapeutic targets and a new class of biomarkers for the cancer prognosis and diagnosis. For example, the lncRNA PCA3 (prostate cancer antigen 3) is a FDA-approved biomarker for prostate cancer prediction. The overexpression of lncRNA HOTAIR in breast cancer patients is reported to be associated with patient survival and risk of metastasis [7]. Another important lncRNA ANRIL (CDKN2-AS1) is one of the most frequently alerted genes in human cancers and has been reported to increase cancer risks in diverse cancers.

Although a large number of lncRNAs have been identified, only a very small proportion of them have been characterized for cellular and molecular functions. Similar to protein-coding genes, the biomarker discovery of lncRNAs can start from a genome-wide differential expression (DE) analysis. One advantage of lncRNAs research in cancer is that we can leverage the large collection of previously published RNA-seq data and perform secondary analyses. Unlike the miRNAs counterparts, the expression of a large number of lncRNAs can be detected by standard RNA-seq with sufficient sequencing depth. Through downloading RNA-seq BAM files and recalling using GENCODE genomic coordinates, more than 8,000 human tumor samples across all major cancer types in The Cancer Genome Atlas (TCGA) and other published studies have been re-analyzed for the lncRNAs expression profile [8, 9]. There is a limited number of non-tumor samples sequenced for RNA-seq in TCGA. If necessary, the database such as the GTEx (http://gtexportal.org) can serve as additional tissue-specific controls, which provides over 9,600 RNA-seq samples across 51 tissues.

lncRNAs expression data have several features that pose significant challenges for the data analysis, including low abundance, large number of genes, and rough annotations. To ensure detection reliability, a common practice is to filter out lncRNA genes with low average Reads Per Kilobase per Million mapped reads (RPKM), e.g. <0.3. We recommend using the two-step filter proposed in [9]: in the first step eliminates genes with 50th-percentile RPKM =0, and in the second step only keep genes whose 90th-percentile RPKM <0.1. About two-thirds of lncRNAs are excluded after this filtering procedure. Interestingly, excess zeros or low expression values are still observed in the downsized dataset. It is well known that excess zero read counts in RNAseq data can distort model estimation and reduce power in differential expression analysis. The popular R packages DESeq2 and edgeR assume a negative

binomial (i.e. over-dispersed Poisson) distribution for the count data. Methods based on zero-inflated negative binomial (ZINB) and zero-inflated GLM have been proposed to explicitly address the issue of excess zeros in RNA-seq data [10]. These methods have been recently applied to single-cell RNA-seq (scRNA-seq) data, which has high dropout rates. Since the difference in gene expression variance is biologically interesting, multiple methods have been developed to incorporate the testing of variance in the differential model. However, for biomarkers in clinical settings, genes with pronounced group contrast in mean expression level usually have more translation value. Gene wise expression variability can generate from different sources and varies widely from study to study, especially with different normalization methods. Hence, we focus on the group comparison of mean gene expression level regardless of variability in this study.

In a large-scale secondary analysis of expression data such as in lncRNA studies, only normalized data (such as RSEM or RPKM) are available [11, 12]. Certain packages such as DESeq2, however, cannot be applied because they do not accept normalized expression and zero as input. In this case, a common practice is to round continuous expression values into integers and shift it to be nonzero. Another commonly-adopted approach is using $log_2\,(x+1)$ transformed normalized data in R package like limma [13], i.e., assuming a log-transformed Gaussian distribution as in microarray intensity levels. The core function in limma, which basically runs a moderated t-test after an empirical Bayes correction, is more generic and more suitable for the differential expression of processed lncRNA expression data. In a very recent study, a total of 25 popular methods for testing differential expression genes were comprehensively evaluated with special emphasis on low-abundance mRNAs and lncRNAs [14]. It was observed that linear modeling with empirical Bayes moderation (implemented in limma with variance stabilizing transformation [15], voom [16] or trend), and a non-parametric method based on Wilcoxon rank sum statistic (implemented in SAMSeq) showed overall good balance of false discovery rate (FDR) and reasonable detection sensitivity. However, none of the methods compared can outperform all other tools and all tools exhibited substandard performance for lncRNAs in terms of differential testing, often with higher FDR and true positive rate (TPR) than for mRNAs. This study also concluded that accurate differential expression inference of lncRNAs requires more samples than that of mRNAs. Even methods like limma can exhibit an excess of false discoveries under specific scenarios, making these methods unreliable in practical applications.

In this paper, we present the lncDIFF, an efficient and reliable toolset based on a zero-inflated exponential quasi-likelihood strategy without the need to fully specify a parametric model. The quasi-likelihood model provides unbiased and efficient estimators even under erroneous assumptions about density. It thus provides a simple and versatile approach to model gene expression data without making strong distributional assumptions about the underlying variation, but still being compatible with existing RNA-Seq quantification and normalization tools. The flexibility in allowing for the estimation of calibration and variance parameters is especially important for lncRNAs differential analysis. The lncDIFF is thus able to integrate desirable features from the aforementioned two top-performing methods (limma and

SAMSeq [14]) for lncRNA differential analysis. The lncDIFF is compared with existing tools using an extensive simulation study and real data analysis on TCGA head and neck squamous cell carcinomas (HNSC). Results suggest that lncDIFF is powerful and robust in a variety of scenarios and identifies DE lncRNA genes of low expression with more accuracy.

METHOD

RNA-Seq Counts Distribution Based Variation

In RNA-Seq gene expression analysis, the type of RNAs and the selected alignment, quantification and normalization tools usually have substantial impact on the distribution pattern of transcript abundance as discussed in [17], especially on the level of gene expression dispersion, i.e. the mean-variance relation. Most of the existing RNA-Seq tools, such as DESeq [18], edgeR [19], and baySeq [20] estimate gene-wise counts dispersion to perform raw counts normalization or differential expression analysis. However, this technique may not be suitable for low-abundance mRNA or lncRNA. The analysis tools such as limma [13, 21] with data transformation become superior for lncRNA instead [14]. In other words, the underlying mean-variance relation distinguishes different types of RNA-Seq counts and determines the tools for downstream analysis.

Let X_{gi} represent RNA-Seq read counts mapped to gene g in sample $i, g = 1, \dots, G, i = 1, \dots, N$. The existing analysis on RNA-Seq data usually assumes Negative Binomial (NB) or the Log Normal (LN) distribution for raw or normalized counts [14, 16], with mean-variance relation summarized as a quadratic form $Var(X_{gi}) = c \cdot E(X_{gi})^2$. The positive constant c is the 'variation' parameter, i.e., the square of coefficient of variation (CV) and depends on the density, i.e. $c = \phi + \frac{1}{\mu_1}$ for NB and $c = \exp(\sigma^2) - 1$ for LN [18]. The parameters μ_1, ϕ are the mean and dispersion of NB, and σ is the log standard deviation of LN, not affected by the log mean.

We use the lncRNA and mRNA data in the TCGA HNSC study to investigate the variation patterns for different types of sequencing counts. If the quantified RNA-Seq reads follow a NB distribution, the gene-wise variation or CV changes inversely with mean expression level, as revealed by the violin box plots for mRNA normalized counts in **Figure 1**. In contrast, the CV level for lncRNA normalized counts in the same study presented by **Figure 1** does not change along with gene-wise mean at 20th-50th percentiles, similar to the LN distribution in which CV is independent of (log) mean.

For mRNA read counts, NB density is a valid assumption and provides robust estimate for the mean expression, dispersion and DE analysis. However, the scale and distribution of lncRNA counts varies across genes, some of which are extremely low and similar to LN, or have a mixture distribution of NB and LN. The differential analysis for a large number of lncRNA genes with mean expression ranging from less than 1 to over 100 can be severely biased, if one uses NB model to account for dispersion

across all genes, or transformation such as log2 [22], voom [16] and variance stabilizing transformation (VST) [15] to remove the skewness for all genes. In the light of fewer statistical assumptions and parameters, it is worth to investigate the plausibility of utilizing Exponential density in RNA-Seq analysis, for which the quadratic mean-variance relation is $\mathrm{CV} = 1$.

Exponential Quasi-Likelihood

In this study, we only consider the normalized lncRNA expression data, i.e. Reads Per Kilobase per Million mapped reads (RPKM) [23] or Fragments Per Kilobase per Million mapped reads (FPKM), as the aim is to improve hypothesis testing of treatment or biological group effect on lncRNA expression regardless of the latent variation pattern. The common normalization methods, such as UQ, TMM [24, 25] are also compatible with lncDIFF, but not assessed in the simulation study and real data analysis, due to limited publically available lncRNA raw counts. We will demonstrate that the choice of normalization method does not affect the validity and accuracy of parameter estimation and DE analysis results in lncDIFF. See the last subsection of Method.

Let Y_{ij} be the lncRNA RPKM for gene i in sample j, belonging to phenotype or treatment group $k, k = 1, \dots, K$. The generalized linear model (GLM) for Y_{ij} with the Exponential family is

$$Y_{ij} \sim \text{Exponential}(\lambda_{ij}), \lambda_{ij} = E(Y_{ij})$$

$$\text{Identity link: } \lambda_{ij} = \sum_{k=1}^{K} \beta_{ik} w_{jk} + \sum_{m=1}^{M} \gamma_m v_{jm}$$

$$\text{Logarithmic link: } \log(\lambda_{ij}) = \sum_{k=1}^{K} \beta_{ik} w_{jk} + \sum_{m=1}^{M} \gamma_m v_{jm}$$

w_{jk} and β_{ik} are design matrix elements and unknown coefficients for groups, v_{jm} and γ_m are the covariates and corresponding coefficients. Since Y_{ij} has been normalized for library size, this model does not include the RNA sequencing normalization factor, although it is a common parameter in existing tools based on NB assumption [18, 19, 26, 27].

In the absence of zero expression, lncDIFF uses the Exponential GLM to lncRNA RPKM DE analysis regardless of the true density of Y_{ij} as a quasi-likelihood approach, which uses a distribution-free statistics to estimate group-wise mean RPKM, similar to the pseudo likelihood (PL) and quasi likelihood (QL) for dispersion estimate in [27]. Let $\beta_i = (\beta_{i1,\dots,}\beta_{iK})$ and $\gamma = (\gamma_{1,\dots,}\gamma_m)$, for gene i with negligible zero occurrence (<1%), the GLM likelihood based on the exponential density $f(Y_{ij}) = \frac{1}{\lambda_{ij}} e^{-\frac{Y_{ij}}{\lambda_{ij}}}$ with identity or log link function is

$$\text{Identity link: } L(\beta_i, \gamma) = \sum_{j=1}^{N} l(\beta_i, \gamma) = \sum_{j=1}^{N} - \left[\frac{Y_{ij}}{\sum_{k=1}^{K} \beta_{ik} w_{jk}} + \log\left(\sum_{k=1}^{K} \beta_{ik} w_{jk} + \sum_{m=1}^{M} \gamma_m v_{jm}\right) \right] \quad (1)$$

Logarithmic link: $L(\beta_i,\gamma)=\sum_{j=1}^{N} l(\beta_i,\gamma) = \sum_{j=1}^{N} -\left[Y_{ij}e^{-(\sum_{k=1}^{K}\beta_{ik}w_{jk})} + \sum_{k=1}^{K}\beta_{ik}w_{jk} + \sum_{m=1}^{M}\gamma_m v_{jm}\right]$ (2)

The exponential quasi likelihood estimate for mean RPKM in lncDIFF is the maximizer of $L(\beta_i,\gamma)$, that is $(\hat{\beta}_i,\hat{\gamma}) = argmax\, L(\beta_i,\gamma)$. This estimate does not require prior knowledge about the statistical distribution of RPKM values, and accommodates the genes with a wide range of expression, i.e. having both extremely low (RPKM<1) and regular (RPKM>10) abundance in a large proportion of samples. The commonly adopted statistical assumptions like Poisson, NB or LN densities about RNA-Seq counts are still allowed in lncDIFF. However, the specified density does not affect the estimation of mean RPKM ($\hat{\beta}_i$) and the corresponding DE analysis results, as illustrated in the Supplementary Methods and simulation study.

Zero-Inflated Exponential Quasi Likelihood

In lncRNA expression data, it is common to observe zero value for a gene in a non-negligible proportion (i.e., at least 1%) of samples. The excess zeroes in lncRNA RPKM cannot be addressed by integer models like Poisson and Negative Binomial (or Gamma-Poisson), since RPKM for most lncRNA genes are non-integer and fall in the range of (0, 2). Rounding decimals to integers and then applying Poisson or NB density [22, 28] or using data transformation, e.g. log2, voom, or VST [15, 16, 22] with limma [13, 21] may lead to errors in DE analysis. Therefore, we propose the zero-inflated quasi likelihood for the GLM of Y_{ij} to account for the inflation of zeros in lncRNA expression.

In order to incorporate the zero-inflated pattern, we first re-specify the RPKM for gene i in sample j by a multiplicative error model [29-31] with random error ϵ_{ij}, that is

$$Y_{ij} = \lambda_{ij}\epsilon_{ij},\ \ E(\epsilon_{ij}) = 1 \quad (3)$$

The random errors ϵ_{ij} also have the occurrence of excess zeros with a prior probability mass $P(\epsilon_{ij}=0) = 1-\pi,\ P(\epsilon_{ij}>0) = \pi$, and a continuous density at positive value with $E(\epsilon_{ij}|Y_{ij}>0) = \gamma$, similar to [30, 32, 33]. If the positive RPKM $Y_{ij}|Y_{ij}>0$ follows the Exponential distribution (so does $\epsilon_{ij}|Y_{ij}>0$), then the density functions for Y_{ij} including zero occurrence is

$$f(Y_{ij}) = (1-\pi)^{I(Y_{ij}=0)}\left(\frac{\pi^2}{\lambda_{ij}}e^{-\pi Y_{ij}/\lambda_{ij}}\right)^{I(Y_{ij}>0)} \quad (4)$$

Equation (4) is derived in the Supplementary Methods. See supplementary data files. Similar to the aforementioned Exponential quasi-likelihood for GLM, lncDIFF applies the zero-inflated density in equation (5) to GLM as a quasi-likelihood approach to perform DE analysis of zero-inflated lncRNA expression. The corresponding quasi-likelihood function is

$$L^*(\pi,\beta_i,\gamma) = \sum_{j=1}^{N} l_j^*(\pi,\beta_i,\gamma) \quad (5)$$

$l_j^*(\pi, \beta_i, \gamma)$ is defined according to the selected link function as

Identity link:

$$l_j^*(\pi, \beta_i, \gamma) = I_{(Y_{ij}=0)} \log(1-\pi) + I_{(Y_{ij}>0)} \left(2 \cdot \log(\pi) - \frac{\pi Y_{ij}}{\sum_{k=1}^{K} \beta_{ik} w_{jk}} - \log\left(\sum_{k=1}^{K} \beta_{ik} w_{jk} + \sum_{m=1}^{M} \gamma_m v_{jm}\right)\right)$$

Logarithmic link: $l_j^*(\pi, \beta_i, \gamma) = I_{(Y_{ij}=0)} \log(1-\pi) + I_{(Y_{ij}>0)} \left(2 \cdot \log(\pi) - \pi Y_{ij} e^{-\left(\sum_{k=1}^{K} \beta_{ik} w_{jk} + \sum_{m=1}^{M} \gamma_{ik} v_{jm}\right)} - \sum_{k=1}^{K} \beta_{ik} w_{jk} - \sum_{m=1}^{M} \gamma_m v_{jm}\right)$

The zero-inflated quasi-maximum likelihood (ZI-QML) estimate for group-wise mean RPKM is the maximizer of $L^*(\pi, \beta_i, \gamma)$ in equation (6), that is

$$(\hat{\pi}, \hat{\beta}_i, \hat{\gamma})_{ZI-QML} = argmax\, L^*(\pi, \beta_i, \gamma) \qquad (6)$$

It is worthwhile to note that the likelihood function $L^*(\pi, \beta_i, \gamma)$ in equation (5) reduces to equations (1) and (2) if the proportion of zero expression is negligible, i.e. no more than 1%.

Likelihood Ratio Test

For differential analysis in lncDIFF, we apply the Likelihood Ratio Test (LRT) to the zero-inflated exponential likelihood function $L^*(\pi, \beta_i, \gamma)$ to test hypothesis: H_0: $\beta_i = \beta_{null}$ vs H_1: $\beta_i = \beta_{full}$, where β_{null} is the design matrix coefficients with some equal to zero and β_{full} is the coefficients without zero. The test statistic of LRT is $D = -2L^*(\beta_{null}) + 2L^*(\beta_{full})$ with β_{null} and β_{full} being the design matrix coefficients for null and alternative models. Let m_{null} and m_{full} be the number of distinct coefficients in β_{null} and β_{full}. Test statistic D asymptotically follows χ^2 distribution with degrees of freedom $m_{full} - m_{null}$. The p-values from LRT are adjusted for multiple testing using the procedure of Benjamin and Hochberg false discovery rate [34]. The choice of link function does not affect the power of LRT, as illustrated by simulation study.

We also provide empirical distribution of LRT statistics D to compute the p-values for DE analysis, similar to [28]. The empirical distribution of statistics D per gene can be generated by randomly shuffling the samples into *K* groups for *P* times and then calculate the LRT statistics for each permutation, that is $D_1, \ldots, D_P$. Let the test statistics for the true groups be D_0, then the empirical p-value is $\frac{\sum_{p=1}^{P} I_{(D_p > D_0)}}{P}$, and can be adjusted by Benjamin and Hochberg procedure. We implemented the LRT for lncRNA DE analysis based on ZI-QML with observed and empirical p-values.

lncDIFF on Other Normalization Methods

lncDIFF adopts the estimator $(\hat{\pi}, \hat{\beta}_i, \hat{\gamma})_{ZI-QML}$ in equation (6) to estimate the mean gene expression level, based on a likelihood function that captures zero-inflation pattern without assuming the true density of non-zero RPKM. We can theoretically prove that this estimate is asymptotically unbiased, i.e., $(\hat{\pi}, \hat{\beta}_i, \hat{\gamma})_{ZI-QML}$ converges to the true value of (π, β_i, γ) as sample size increases.

According to [35, 36], $(\hat{\pi}, \hat{\beta}_i, \hat{\gamma})_{ZI-QML}$ is asymptotically unbiased as long as $L^*(\pi, \beta_i, \gamma)$ converges to $E[l_j{}^*(\pi, \beta_i, \gamma)]$ and $E[l_j{}^*(\pi, \beta_i, \gamma)]$ is uniquely maximized at the true value, i.e. $\pi_0, \beta_{i0}, \gamma_0$. Suppose $\beta_{i0} = (\beta_{i10}, \dots, \beta_{iK0})$, $\gamma_0 = (\gamma_{10,\dots,}\gamma_{m0})$, for identity link function, the true expectation of Y_{ij} is $\lambda_{ij0} = \sum_{k=1}^{K} \beta_{ik0} w_{jk} + \sum_{m=1}^{M} \gamma_{m0} v_{jm0}$. By law of large numbers, it is not hard to show that $L^*(\pi, \beta_i, \gamma)$ converges to $E[l_j{}^*(\pi, \beta_i, \gamma)]$, where

$$E[l_j{}^*(\pi, \beta_i, \gamma)] = E[l_j{}^*(\pi, \lambda_{ij})] = (1 - \pi_0)\log(1 - \pi) + \pi_0 \left(2 \cdot \log(\pi) - \frac{\pi \lambda_{ij0}}{\pi_0 \lambda_{ij}} - \log(\lambda_{ij})\right)$$

$E[l_j{}^*(\pi, \beta_i, \gamma)$ being uniquely maximized at $(\pi_0, \beta_{i0}, \gamma_0)$ is demonstrated by maximizing the term $E[l_j{}^*(\pi, \lambda_{ij}, \gamma)] + \pi_0 \log(\lambda_{ij0})$, which does not depend on the distribution assumption about non-zero lncRNA RPKM. Detailed proof for unbiased estimate in either link function is elaborated in the Supplementary Methods. The use of Exponential family quasi likelihood in lncDIFF guarantees the accuracy of mean expression estimate under unknown distribution of RPKM Y_{ij}. Thus, lncDIFF mean expression estimation on other normalized counts, (i.e., normalized by TMM or UQ) is also asymptotically unbiased, as long as the normalized counts are non-negative.

In order to illustrate normalization method having no impact on lncDIFF performance, we simply applied lncDIFF DE analysis to three different types of normalized counts (i.e., FPKM, TMM and UQ) of low abundance mRNA in TCGA HNSC tumor-normal samples (N=546). The low abundance genes are selected with mean FPKM in the range of (0.3, 2) and no more than 20% zero expression, similar to the majority of lncRNA genes. The Pearson correlation of log10 adjusted p-values between the three normalization methods are FPKM vs TMM 0.82, FPKM vs UQ 0.92, TMM vs UQ 0.96, implying similar DE analysis results. Therefore, we only use RPKM of lncRNA in TCGA HNSC study to illustrate the application and performance of lncDIFF in Results.

In addition to TMM and UQ, the distribution-free parameter estimation and LRT in lncDIFF are also compatible with model-based RNA-Seq quantification and normalization tools, such as RSEM [37], baySeq [20], and QuasiSeq [38]. Hence, the lncDIFF DE analysis can be incorporated into existing RNA-Seq quantification and normalization pipeline, regardless of the models employed in the preprocessing tools.

RESULTS

Simulation study to assess lncDIFF performance

We conducted a comprehensive simulation study to demonstrate the performance of ZI-QML with observed p-value of LRT and compare to existing common tools DESeq2, edgeR and limma (with log transformation). We rounded decimals to integers as input for DESeq2 and selected the quasi-likelihood estimation method in edgeR. The testing methods for DE genes were LRT in DESeq2 and edgeR, F-test in limma. We considered NB and LN as true densities for data sampling, and used the gene-wise estimate for dispersion or log variance from a real lncRNA RPKM dataset to determine the values of ϕ (NB) and σ^2 (LN) in data generating functions. Based on the dispersion and log variance estimate for the data in TCGA head and neck squamous cell carcinomas (HNSC) study [39], we adopted $\phi = 1, 2, 10, 20$, $\sigma^2 = 0.01, 0.25, 1, 2.25$, and then used fixed ϕ, σ^2 values to generate RPKM of each genes across all samples in the same simulation scenario. Each scenario is defined by the unique gene-wise nonzero proportion $\pi = 0.5,\ 0.7,\ 0.9, 1$, sampling density function (NB or LN) and value of ϕ, σ^2, with sample size varying at N=100, 200, 300.

In order to generate data similar to lncRNA RPKM, we first obtain binary outcomes (0-1) for all samples in one scenario from the Bernoulli sampling, and then replace the 1's by positive values generated by NB or LN densities. It should be noted that in our model with identity link function, the expectation of non-zero RPKM per gene per sample is $E(Y_{ij}|Y_{ij} > 0) = \lambda_{ij}\gamma = \frac{\lambda_{ij}}{\pi}$. Hence, the non-zero RPKM for gene i in group k are randomly generated from NB or LN densities with mean at $\frac{\beta_{ik}}{\pi}$, where β_{ik} being the mean of gene i (including zero expression) in group k. The HNSC study includes 40 pairs of matched normal-tumor tissues. We use the 40 normal samples to calculate the mean RPKM as baseline group parameter β_{i1} in simulation. Similar to the common filtering criteria in existing lncRNA analysis, we remove the genes in the real data with mean RPKM <0.3 [40, 41] and zero expression in more than half of the samples, reducing to 1100 genes for simulation.

In the simulation study, we only considered two-group comparison to illustrate the contrast between different methods. RPKM of the first group is randomly-generated by the specified density function and the baseline parameter, while the second group has the mean parameter of the baseline times a shift, i.e., the tumor/normal fold change in TCGA HNSC data. We manually set the shift between two simulated groups at 1 if the absolute log2 fold change (LFC) for the corresponding gene is less than 0.5. Simulated genes with between-groups shift at 1 are the null genes and the remaining are DE genes. For each simulated scenario, we generate 100 replicates to assess the performance of different methods by the mean of Type I error, false discovery rate (FDR), true positive rate (TPR), and area under the curve (AUC) of receiver operating characteristics (ROC) with FDR threshold 0.05.

We order the scenarios by the level of variance (with 1-4 representing the smallest to the largest and determined by dispersion or log variance), proportion of nonzero expression, and sample size to investigate the impact of parameters on performance metrics. **Figure 2** and **Supplementary Figures S1-**

S3 present the AUC, FDR, TPR and Type I error (or false positive rate) of all scenarios, illustrating that ZI-QML outperforms the other methods, especially for scenarios with LN density. AUC for all methods in **Figure 2** decrease as the gene-wise variation increases, and it also shows that ZI-QML's performance is close to the optimal method (DESeq2) for NB density. The change of AUC across different sample sizes implies that adding more samples improves the performance of ZI-QML and DESeq2, but does not have impact on edgeR and limma. Furthermore, the AUC of ZI-QML in NB density is equivalent to or slightly larger than that of DESeq2 at sample size N=400. According to AUC and TPR, the outperformance of DESeq2 compared to lncDIFF for NB density is not as pronounced as the outperformance of ZI-QML compared to DESeq2 for LN density. On the other hand, the FDR and Type I error show that lncDIFF has similar performance of DESeq2 in most scenarios regardless of density and greatly outperforms the other two methods, although lncDIFF in large-variance LN scenarios presents performance close to edgeR and limma. In summary, lncDIFF is the optimal method for DE analysis of lncRNA RPKM with different distributions, and DESeq2 is an ideal tool if the non-zero abundance is relatively high and follow NB density.

Application of lncDIFF to TCGA HNSC Data

We first employed the same methods to perform DE analysis on the TCGA HNSC lncRNA data for matched tumor and normal samples. The Venn diagram in **Figure 3 (A)** shows the overlap and difference of the DE genes identified by four methods. In real data analysis, the proportion of genes with absolute log2 fold change >0.5 and identified as DE is an alternative metric for the true positive rate, while the false positive rate (FPR) can be approximated by the proportion of genes with absolute LFC less than 0.5, 1, 1.5 but identified as DE. The significance threshold for tumor vs normal is set at FDR<0.05. We listed the alternative TPR and FPR in **Figure 3 (B)**, and presented the contrast between lncDIFF and the other methods by boxplots in **Figure 3 (C)-(E)**, with each panel showing the tumor vs normal group effect on the lncDIFF positive genes identified as negative by other methods. We only include the genes with upregulation for normal tissues and LFC>0.5 in the boxplots. The results in **Figure 3 (B)** confirms that lncDIFF provides ideal power or alternative TPR (75%) in DE analysis for LFC<0.5, with approximated FPR below 5%. The other methods either have TPR no more than 30% or generates false positives with an approximate probability 44%. The boxplots in **Figure 3 (C)-(E)** reveal that the group contrast on DE genes identified only by lncDIFF is larger than that identified only by the other methods. This also implies that lncDIFF is less likely to 'miss' the DE genes with large group contrast.

We also applied the same analysis to the unpaired tumor (N=426) vs normal (N=40) samples in the TCGA HNSC study by lncDIFF, and compared the top significant genes in the paired and unpaired DE analysis results, listed in **Table 1**. There are 11 overlapped genes in the top-20 significant gene list of paired and unpaired analysis, some of which are associated with overall survival time. For each the overlapped significant genes, we divided the 426 HNSC tumor samples into two groups by the median of

RPKM per DE gene, and then apply Cox Proportional Hazard model to survival association analysis. The Kaplan-Meier curves and the log-rank test p-values reveal marginal or significant association between genes *ERVH48-1*, *HCG22, LINC00668, LINC02582* and the overall survival months, illustrated by **Figure 4**. For the same set of HNSC tumor samples, we also used the mRNA RSEM normalized counts to select 20 mRNA genes highly correlated with the 11 tumor-normal DE lncRNA genes by Spearman correlation, listed in the **Supplementary Table 1**.

Secondly, we used 72 TCGA HNSC tumor samples with valid Human Papillomavirus (HPV) status (i.e. positive vs negative) to compare DE analysis results by different methods. The FDR threshold is set at <0.1 for this analysis, since the contrast between HPV positive and negative is less pronounced compared to tumor vs normal. The Venn diagram and table in **Supplementary Figure S4 (A)-(B)** show the overlap of DE genes between different methods along with the approximate power and FPR. lncDIFF still provides more power with FPR controlled at 0.02, while the other methods have DE analysis power close to either zero or FPR. **Supplementary Figure S4 (C)** are the PCA plots generated by the top 200 significant genes in terms of the FDR of each method. Based on the distance between clusters, lncDIFF top significant genes differentiate the HPV status better than those identified by the other methods do.

DISCUSSION

lncDIFF is an efficient and powerful differential analysis tool for lncRNA RPKM or FPKM. The distribution of lncRNA RPKM is different from that of mRNA, as some genes in lncRNA may have low or even zero expression for a subset of samples, but also have normal expression for the remaining samples. Existing RNA sequencing analysis tools based on a unique density assumption ignore such characteristic and does not take excess of zeros into account. For example, DESeq2 does not allow zero counts and decimals as input data; hence, RPKM must be rounded and transformed to nonzero integers, which reduces the variation of low abundance genes across samples. Although edgeR handles non-integer RNA-Seq expression data and allows zero values, the computation for group effect depends on the estimate of gene-wise dispersion, which can be severely biased for a gene having both normal and low expression occurrence.

The Exponential likelihood function used in lncDIFF is not derived from the true density of lncRNA RPKM, but the group effect estimate based on that is valid and asymptotically unbiased, as demonstrated by the proof in Supplementary Methods. It is worth to note that this result does not hold for Poisson, NB, or LN likelihood function if the gene expression density is incorrectly assumed, since the demonstration is based on the unique structure of Exponential density function and not applicable to other distribution families. The choice of link function does not have any impact on the group effect estimate and LRT results as shown by **Table 2**, but the log link function can avoid NA values produced in numerical optimization of the likelihood function.

The distribution of p-values from lncDIFF is also investigated and compared with the other methods in TCGA HNSC tumor vs normal analysis, using simulated p-values from sample permutation. We randomly selected three genes with different RPKM density patterns to generate the null p-values and then visualized the p-values distribution via QQ plots in **Figure 5**. **Figure 5** (B) shows that the p-values of lncDIFF and DESeq2 are similar and close to the expected distribution, while edgeR and limma tend to provide a large proportion of small p-values (<0.1). The histogram and density plot of RPKM presented in **Figure 5** (A) imply that the null p-values of lncDIFF and DESeq2 follow the uniform distribution for normally or highly expressed lncRNA genes (ENSG00000130600.11), and may deviate from the theoretical distribution for low abundance genes (ENSG00000152931.7, ENSG00000153363.8).

We implemented ZI-QML and LRT with either link function in lncDIFF, along with an option of simulated p-values and FDR generated from permutations. This package allows the input expression matrix to be either continuous or discrete and requires group or phenotype factor provided in design matrix format. This package does not contain raw counts normalization functions, but is compatible with non-negative normalized counts from existing RNA-Seq analysis tools. The group effect estimation is implemented in R function ZIQML.fit, separated from likelihood ratio testing included in function ZIQML.LRT.

We also illustrated the computation efficiency of lncDIFF by running on the TCGA HNSC matched tumor-normal samples with ~1130 filtered genes. The processing time (in seconds) of this biological data analysis by lncDIFF, DESeq2, edgeR and limma are 3.17, 4.31, 3.37 and 0.02, respectively. If the option of simulated p-value is enabled, the running time of lncDIFF on this real dataset is prolonged to 267.86 seconds for default 100 permutations, but the correlation between observed and simulated p-values or FDR's is around 0.9. In future study, we will extend the lncDIFF to account for technical excess zeros from biological zeros [10], as well as apply it to lncRNA counts normalized by other methods, such as UQ and TMM.

CONCLUSION

lncDIFF is a novel method utilizing GLM with a distribution free estimator and LRT in differential analysis of lncRNA normalized counts. This is an efficient DE analysis method, being compatible with various RNA-Seq quantification and normalization tools.

ACKNOWLEDGEMENT

This work was supported in part by the Environmental Determinants of Diabetes in the Young (TEDDY) study, funded by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK).

FUNDING

This work was supported in part by Institutional Research Grant number 14-189-19 from the American Cancer Society, and the National Cancer Institute, part of the National Institutes of Health under grant number [P50 CA168536], Moffitt Skin Cancer SPORE.

CONFLICT OF INTEREST

None declared

REFERENCES

1. Kapranov, P., et al., *RNA maps reveal new RNA classes and a possible function for pervasive transcription.* Science, 2007. **316**(5830): p. 1484-1488.
2. Batista, P.J. and H.Y. Chang, *Long noncoding RNAs: cellular address codes in development and disease.* Cell, 2013. **152**(6): p. 1298-1307.
3. Guttman, M. and J.L. Rinn, *Modular regulatory principles of large non-coding RNAs.* Nature, 2012. **482**(7385): p. 339.
4. Ulitsky, I. and D.P. Bartel, *lincRNAs: genomics, evolution, and mechanisms.* Cell, 2013. **154**(1): p. 26-46.
5. Huarte, M., *The emerging role of lncRNAs in cancer.* Nature medicine, 2015. **21**(11): p. 1253.
6. Chaudhary, R. and A. Lal, *Long noncoding RNAs in the p53 network.* Wiley Interdisciplinary Reviews: RNA, 2017. **8**(3): p. e1410.
7. Gupta, R.A., et al., *Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis.* Nature, 2010. **464**(7291): p. 1071.
8. Li, J., et al., *TANRIC: an interactive open platform to explore the function of lncRNAs in cancer.* Cancer research, 2015: p. canres. 0273.2015.
9. Yan, X., et al., *Comprehensive genomic characterization of long non-coding RNAs across human cancers.* Cancer cell, 2015. **28**(4): p. 529-540.
10. Ran, D. and Z.J. Daye, *Gene expression variability and the analysis of large-scale RNA-seq studies with the MDSeq.* Nucleic acids research, 2017. **45**(13): p. e127-e127.
11. Zhang, W., et al., *Comparison of RNA-seq and microarray-based models for clinical endpoint prediction.* Genome Biology, 2015. **16**(1): p. 133.
12. Bouckenheimer, J., et al., *Differential long non-coding RNA expression profiles in human oocytes and cumulus cells.* Scientific Reports, 2018. **8**(1): p. 2202.
13. Ritchie, M.E., et al., *limma powers differential expression analyses for RNA-sequencing and microarray studies.* Nucleic Acids Research, 2015. **43**(7): p. e47-e47.
14. Assefa, A.T., et al., *Differential gene expression analysis tools exhibit substandard performance for long non-coding RNA-sequencing data.* Genome Biology, 2018. **19**(1): p. 96.
15. Soneson, C. and M. Delorenzi, *A comparison of methods for differential expression analysis of RNA-seq data.* BMC Bioinformatics, 2013. **14**(1): p. 91.

16. Law, C.W., et al., *voom: precision weights unlock linear model analysis tools for RNA-seq read counts.* Genome Biology, 2014. **15**(2): p. R29.
17. Li, P., et al., *Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data.* BMC Bioinformatics, 2015. **16**(1): p. 347.
18. Anders, S. and W. Huber, *Differential expression analysis for sequence count data.* Genome Biology, 2010. **11**(10): p. R106-R106.
19. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-140.
20. Hardcastle, T.J. and K.A. Kelly, *baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data.* BMC Bioinformatics, 2010. **11**(1): p. 422.
21. Smyth, G.K., *Limma: linear models for microarray data*, in *Bioinformatics and computational biology solutions using R and Bioconductor*. 2005, Springer. p. 397-420.
22. Assefa, A.T., et al., *Differential gene expression analysis tools exhibit substandard performance for long non-coding RNA–sequencing data.* bioRxiv, 2017.
23. Mortazavi, A., et al., *Mapping and quantifying mammalian transcriptomes by RNA-Seq.* Nature methods, 2008. **5**(7): p. 621.
24. Robinson, M.D. and A. Oshlack, *A scaling normalization method for differential expression analysis of RNA-seq data.* Genome Biology, 2010. **11**(3): p. R25.
25. Bullard, J.H., et al., *Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments.* BMC Bioinformatics, 2010. **11**: p. 94-94.
26. León-Novelo, L., C. Fuentes, and S. Emerson, *Marginal likelihood estimation of negative binomial parameters with applications to RNA-seq data.* Biostatistics, 2017. **18**(4): p. 637-650.
27. Robinson, M.D. and G.K. Smyth, *Small-sample estimation of negative binomial dispersion, with applications to SAGE data.* Biostatistics, 2008. **9**(2): p. 321-332.
28. Chu, C., et al., *deGPS is a powerful tool for detecting differential expression in RNA-sequencing studies.* BMC Genomics, 2015. **16**(1): p. 455.
29. Brownlees, C.T., F. Cipollini, and G.M. Gallo, *Multiplicative error models.* 2011.
30. Hautsch, N., *Capturing common components in high-frequency financial time series: A multivariate stochastic multiplicative error model.* Journal of Economic Dynamics and Control, 2008. **32**(12): p. 3978-4015.
31. A., M.T., *Predicting and Correcting Bias Caused by Measurement Error in Line Transect Sampling Using Multiplicative Error Models.* Biometrics, 2004. **60**(3): p. 757-763.
32. Pierson, E. and C. Yau, *ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis.* Genome biology, 2015. **16**(1): p. 241.
33. Wu, Z., et al., *Two-phase differential expression analysis for single cell RNA-seq.* Bioinformatics, 2018: p. bty329-bty329.
34. Benjamini, Y. and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* Journal of the royal statistical society. Series B (Methodological), 1995: p. 289-300.
35. Amemiya, T. and H.U. Press, *Advanced Econometrics*. 1985: Harvard University Press.
36. Gourieroux, C., A. Monfort, and A. Trognon, *Pseudo Maximum Likelihood Methods: Theory.* Econometrica, 1984. **52**(3): p. 681-700.
37. Li, B. and C.N. Dewey, *RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.* BMC Bioinformatics, 2011. **12**(1): p. 323.
38. Lund Steven, P., et al., *Detecting Differential Expression in RNA-sequence Data Using Quasi-likelihood with Shrunken Dispersion Estimates*, in *Statistical Applications in Genetics and Molecular Biology*. 2012.

39. The Cancer Genome Atlas, N., *Comprehensive genomic characterization of head and neck squamous cell carcinomas.* Nature, 2015. **517**: p. 576.
40. Tsoi, L.C., et al., *Analysis of long non-coding RNAs highlights tissue-specific expression patterns and epigenetic profiles in normal and psoriatic skin.* Genome Biology, 2015. **16**(1): p. 24.
41. Tang, Z., et al., *Comprehensive analysis of long non-coding RNAs highlights their spatio-temporal expression patterns and evolutional conservation in Sus scrofa.* Scientific Reports, 2017. **7**: p. 43166.

TABLES AND FIGURES LEGENDS

Tables

Table 1: lncDIFF output for the top 20 significant genes for paired and unpaired tumor vs normal differential analysis for TCGA HNSC study. The overlap of genes are in bold. Likelihood Ratio Test statistics, p-value and FDR are output from lncDIFF.

Paired Tumor vs Normal					Unpaired Tumor vs Normal				
Gene	Ensembl ID	Log2 Fold Change	Statistics	FDR	Gene	Ensembl ID	Log2 Fold Change	Statistics	FDR
RVH48-1	**ENSG00000233056.1**	0.415	211.767	7.48E-45	***HCG22***	**ENSG00000228789.2**	-2.979	674.029	1.76E-145
VC02487	**ENSG00000203688.4**	-3.747	200.441	1.11E-42	***LINC02487***	**ENSG00000203688.4**	-3.470	625.994	2.46E-135
HCG22	**ENSG00000228789.2**	-3.138	151.425	3.73E-32	*MYHAS*	ENSG00000272975.1	-0.935	324.216	7.70E-70
VC00668	**ENSG00000265933.1**	2.189	148.534	1.20E-31	*LINC01405*	ENSG00000185847.3	-1.366	276.487	1.45E-59
IC02582	**ENSG00000261780.2**	1.027	144.294	8.10E-31	*FALEC*	ENSG00000228126.1	-1.721	252.647	1.82E-54
VC00941	**ENSG00000235884.2**	2.450	138.020	1.59E-29	*TMEM238L*	ENSG00000263429.3	-2.250	235.559	8.06E-51
VC00942	ENSG00000249628.2	1.105	128.195	1.92E-27	*AC005392.2*	ENSG00000231412.2	-2.342	198.936	6.73E-43
VC01234	ENSG00000249550.2	1.755	121.173	5.79E-26	*AC140479.4*	ENSG00000261760.2	-1.471	188.314	1.23E-40
VC02154	ENSG00000235385.1	2.099	120.529	7.12E-26	***ERVH48-1***	**ENSG00000233056.1**	0.444	185.507	4.47E-40
134312.5	**ENSG00000261327.3**	2.064	115.828	6.85E-25	*AC091563.1*	ENSG00000254343.2	-2.185	174.352	1.10E-37
365181.2	**ENSG00000272068.1**	1.191	111.605	5.24E-24	***LINC02582***	**ENSG00000261780.2**	1.009	161.008	8.19E-35
UXAP9	**ENSG00000225210.5**	2.868	110.895	6.87E-24	***LINC00668***	**ENSG00000265933.1**	1.626	154.270	2.23E-33
UXAP8	**ENSG00000206195.6**	2.422	105.798	8.30E-23	*ACBD3-AS1*	ENSG00000234478.1	-1.733	150.782	1.08E-32
FTA1P	ENSG00000225383.2	1.676	103.239	2.80E-22	***LINC00941***	**ENSG00000235884.2**	2.090	150.692	1.08E-32
010343.3	ENSG00000250697.1	1.838	101.397	6.63E-22	***AC134312.5***	**ENSG00000261327.3**	2.146	150.725	1.08E-32
FN1-AS1	ENSG00000236081.1	1.590	101.238	6.74E-22	***DUXAP9***	**ENSG00000225210.5**	2.711	141.890	8.49E-31
VC00520	ENSG00000258791.3	1.570	98.359	2.71E-21	*ABHD11*	ENSG00000225969.1	-1.730	140.148	1.92E-30
134312.2	**ENSG00000260162.2**	1.912	98.157	2.84E-21	***AL365181.2***	**ENSG00000272068.1**	1.008	138.932	3.35E-30
114956.2	ENSG00000248554.1	3.038	96.948	4.95E-21	***DUXAP8***	**ENSG00000206195.6**	2.230	134.501	2.95E-29
CASC9	ENSG00000249395.2	4.019	91.046	9.28E-20	***AC134312.2***	**ENSG00000260162.2**	1.982	129.028	4.42E-28

Table 2: Group effect estimates and likelihood ratio test results of TCGA HNSC tumor vs normal with logarithmic and identity link functions in lncDIFF.

	Logarithmic link function				
Genes Ensembl ID	$\exp(\beta_1)$ (tumor)	$\exp(\beta_2 - \beta_1)$ (contrast)	$\exp(\beta_2)$ (normal)	p-value	FDR
ENSG00000005206.12	0.247	0.811	0.200	0.348	0.528
ENSG00000100181.17	0.737	0.993	0.732	0.974	0.982
ENSG00000126005.11	7.161	1.263	9.043	0.297	0.474
ENSG00000130600.11	181.885	1.571	285.661	0.044	0.115
ENSG00000131484.3	0.362	1.044	0.378	0.846	0.916
	Identity link function				
Genes Ensembl ID	β_1 (tumor)	$\beta_2 - \beta_1$ (contrast)	β_2 (normal)	p-value	FDR
ENSG00000005206.12	0.247	-0.047	0.200	0.348	0.528
ENSG00000100181.17	0.737	-0.005	0.732	0.974	0.982
ENSG00000126005.11	7.160	1.887	9.047	0.297	0.474
ENSG00000130600.11	181.852	103.833	285.684	0.044	0.115
ENSG00000131484.3	0.362	0.016	0.378	0.846	0.916

Figures

Figure1: Violin and box plots for gene-wise coefficient of variation (CV) based on normalized counts of mRNA and lncRNA. Genes are divided into ten groups by the percentile of mean normalized counts. The first panel is plotted on the log2 scaled normalized counts, while the data used in the second panel is not log transformed.

Figure 2: AUC of ROC curve for DE analysis on simulated data. Scenarios are in the order of true density, proportion of non-zero expression values, variance level. The labels 'Variance1'-'Variance4'represent gene-wise variance levels from the smallest to the largest.

Figure 3: Performance of lncDIFF, DESeq2, edgeR and limma on TCGA HNSC matched tumor-normal samples. (A) is the Venn diagram for DE genes identified by each method. (B) lists the proportion of genes with LFC greater or less than 0.5, 1.0, 1.5 being identified as DE by each method. (C)-(E) are the boxplots of log2 RPKM per gene for tumor vs normal. The genes in (C)-(E) are upregulated in normal tissue and LFC>0.5.

Figure 4: Survival time association with DE genes identified in both paired and unpaired TCGA HNSC tumor vs normal analysis. The 426 tumor samples are divided into two groups by the median of RPKM per gene. (A)-(D) are the Kaplan-Meier survival curves for genes *ERVH48-1*, *LINC00668, HCG22, LINC02582 individually.*

Figure 5: QQ plots of simulated null p-values for genes with different RPKM distributions in TCGA HNSC matched samples. (A) presents the histogram and density plot of RPKM for each genes. (B) shows the corresponding QQ plot of null p-values simulated by shuffling the samples.

Figure 1

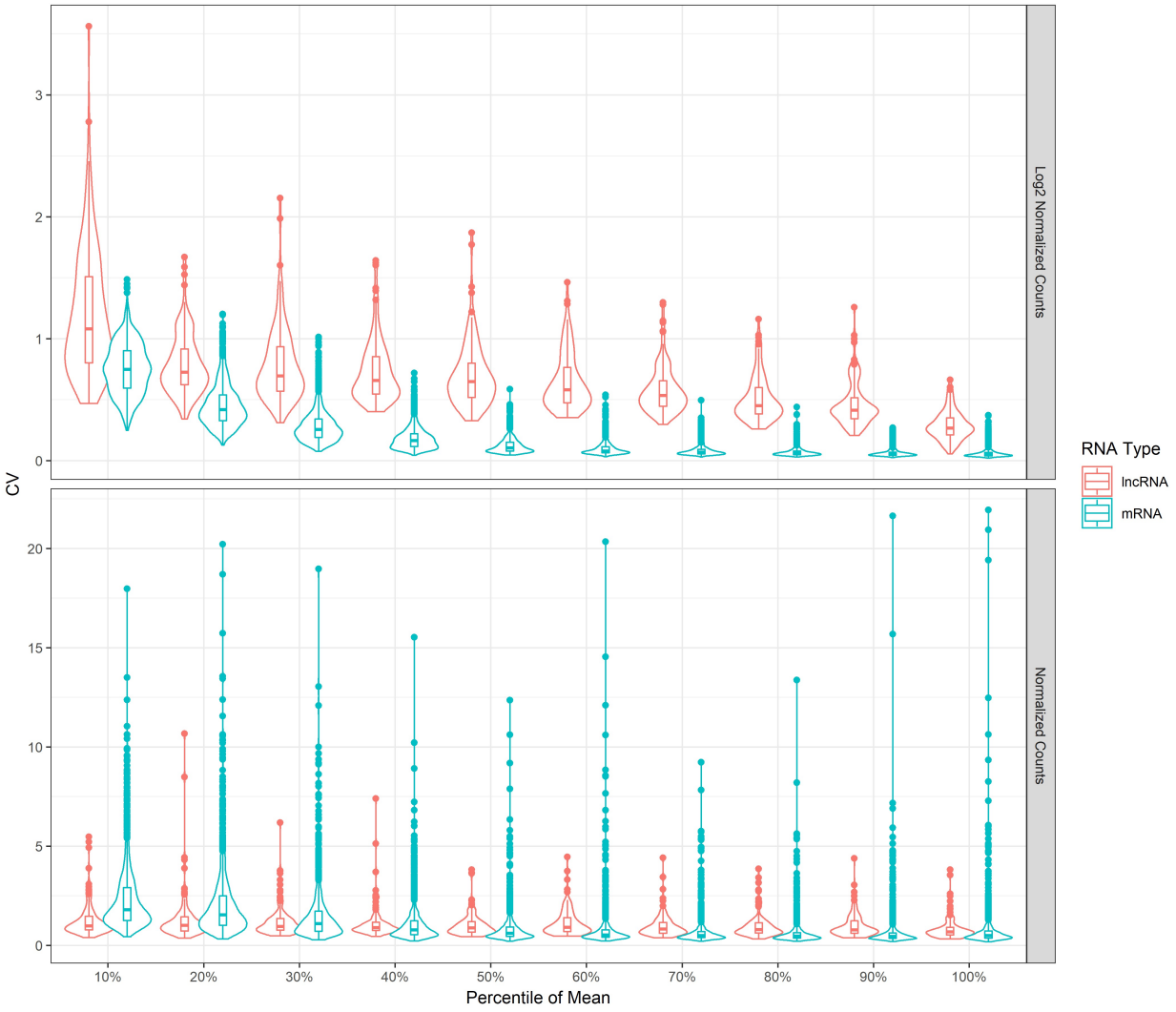


Figure 2

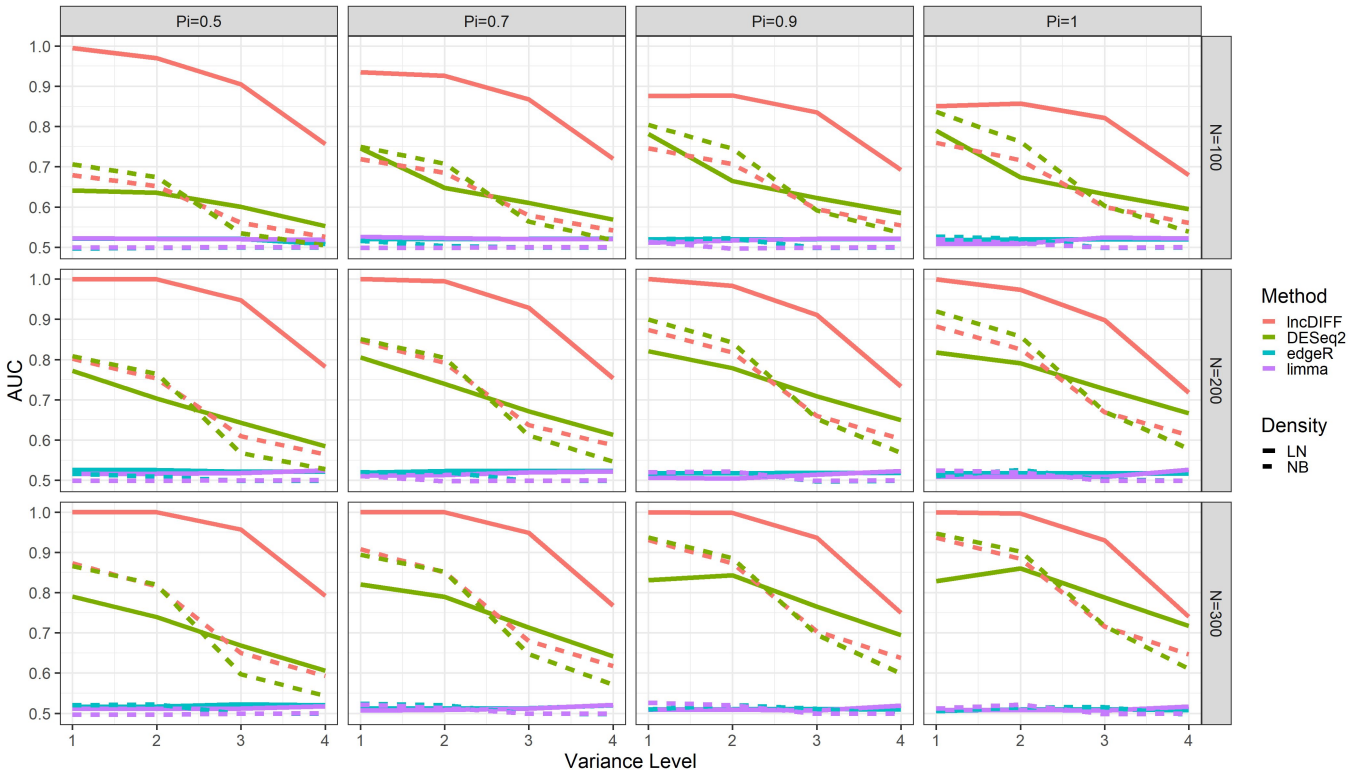
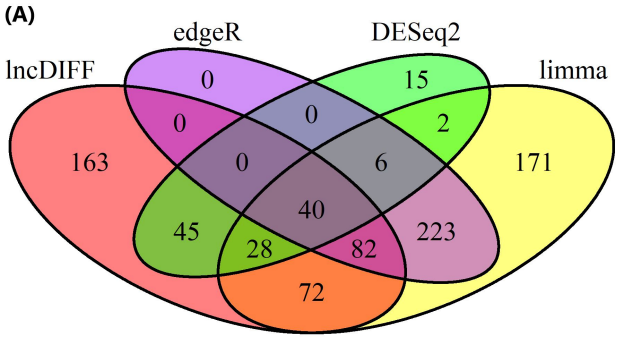


Figure 3



(B)

	lncDIFF	DESeq2	edgeR	limma
LFC>0.5	0.753	0.241	0.297	0.519
LFC<0.5	0.046	0.012	0.245	0.441
LFC>1.0	0.98	0.468	0.298	0.527
LFC<1.0	0.205	0.036	0.26	0.462
LFC>1.5	0.99	0.641	0.34	0.631
LFC<1.5	0.269	0.057	0.259	0.459

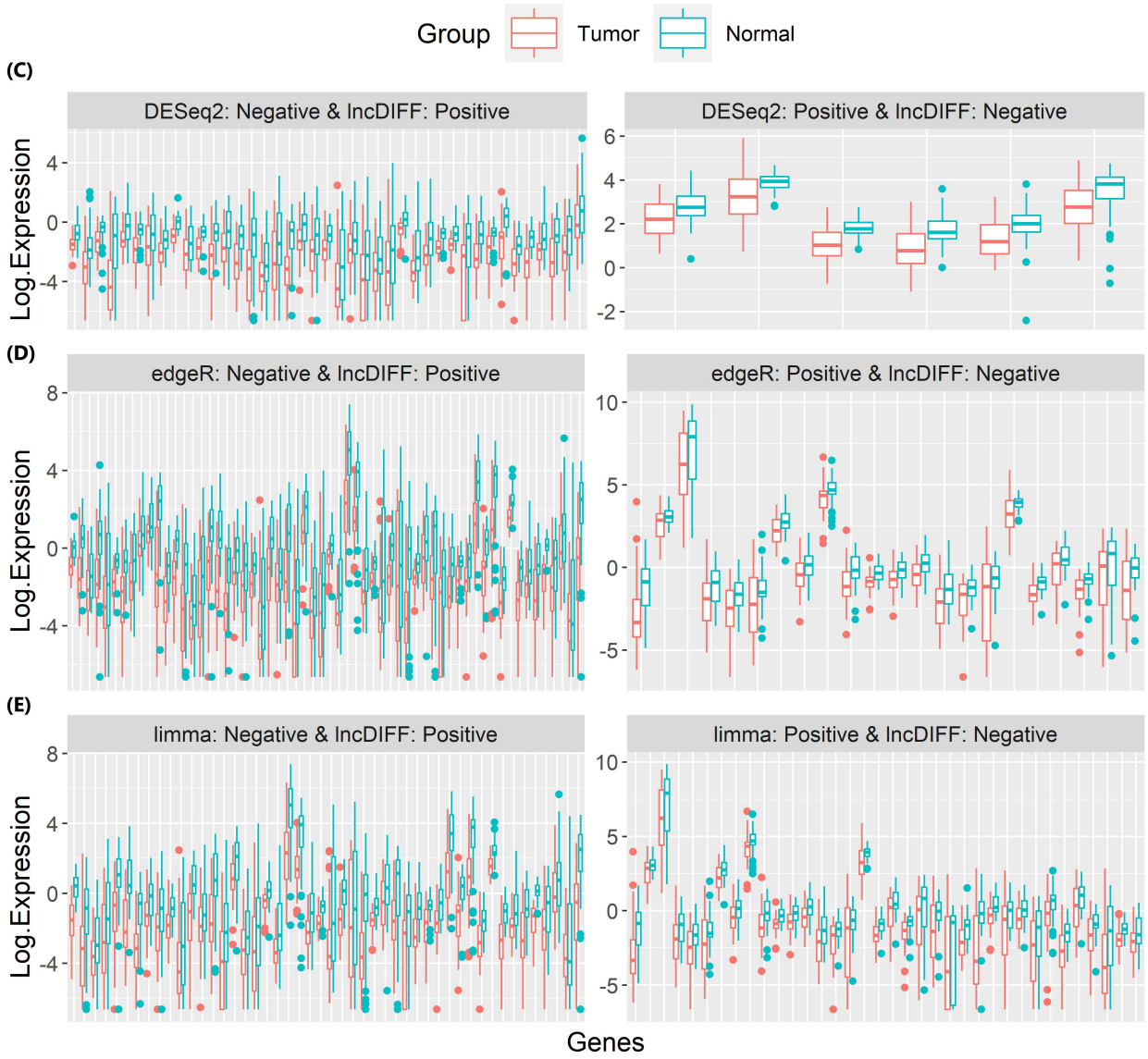


Figure 4

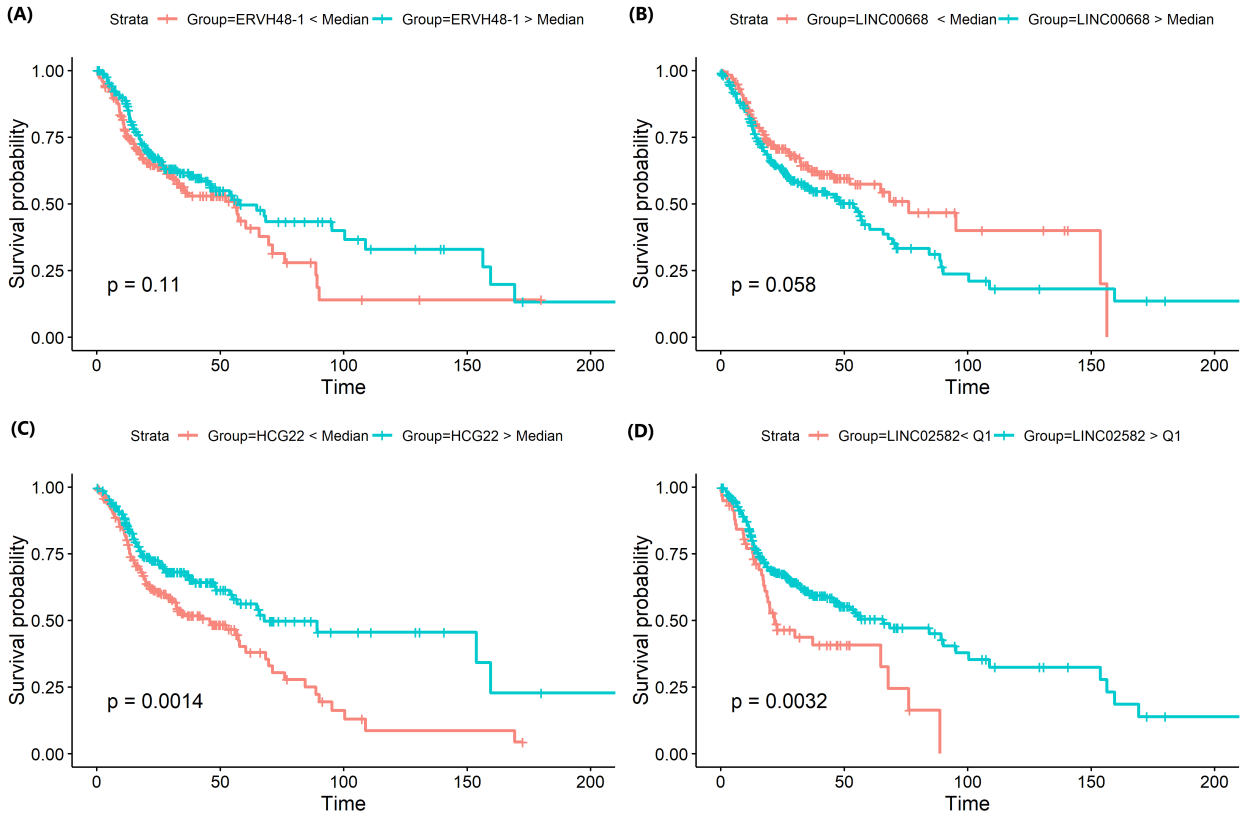
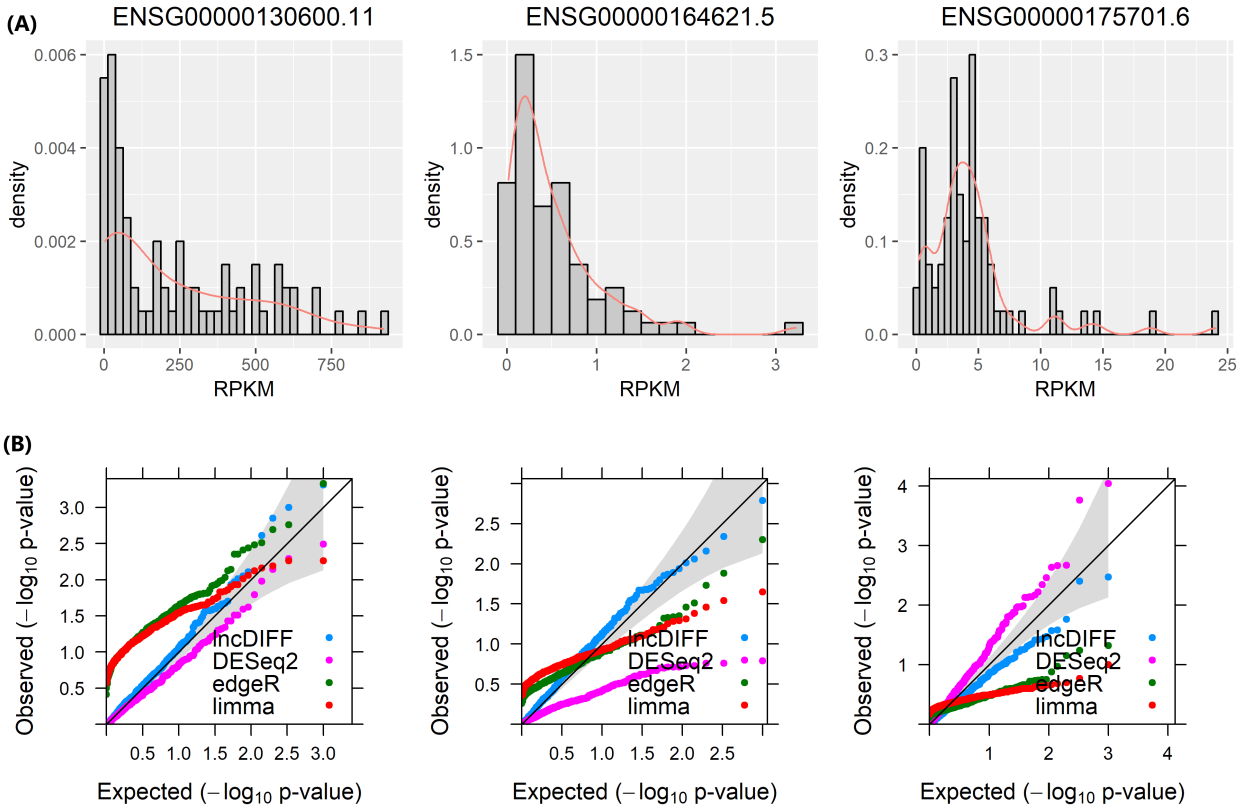


Figure 5



Supplementary methods

1. Zero-Inflated Exponential density for RPKM Y_{ij}

For the multiplicative error model specified in equation (3), if the positive random error $\epsilon_{ij}|Y_{ij} > 0$ follows a distribution described by an Exponential density function $h(\epsilon_{ij}|Y_{ij} > 0) = \frac{1}{\gamma} e^{-\frac{\epsilon_{ij}}{\gamma}}$, then the distribution of ϵ_{ij} including zero occurrence is

$$g(\epsilon_{ij}) = (1-\pi)^{I(\epsilon_{ij}=0)} \left(\frac{\pi}{\gamma} e^{-\frac{\epsilon_{ij}}{\gamma}}\right)^{I(\epsilon_{ij}>0)}$$

with $E(\epsilon_{ij}) = \pi\gamma$. According to the unit mean assumption $E(\epsilon_{ij}) = 1$ in equation (3), we have $\gamma = \frac{1}{\pi}$. That is,

$$g(\epsilon_{ij}) = (1-\pi)^{I(\epsilon_{ij}=0)} (\pi^2 e^{-\pi\epsilon_{ij}})^{I(\epsilon_{ij}>0)}$$

Since $\epsilon_{ij} = Y_{ij}/\lambda_{ij}$, the semi-continuous distribution for Y_{ij} can be derived by

$$f(Y_{ij}) = g(Y_{ij}/\lambda_{ij}) \cdot \frac{d\epsilon_{ij}}{dY_{ij}}$$

That is, $f(Y_{ij}) = (1-\pi)^{I(Y_{ij}=0)} (\frac{\pi^2}{\lambda_{ij}} e^{-\pi Y_{ij}/\lambda_{ij}})^{I(Y_{ij}>0)}$.

2. ZI-QML estimate $(\hat{\pi}, \hat{\boldsymbol{\beta}}_i, \hat{\boldsymbol{\gamma}})_{ZI-QML}$ is asymptotically unbiased.

Proof: According to [1, 2], $(\hat{\pi}, \hat{\beta}_i, \hat{\gamma})_{ZI-QML}$ is a consistent estimator if $L^*(\pi, \beta_i, \gamma)$ converges almost surely to $E[l_j{}^*(\pi, \beta_i, \gamma)]$ and $E[l_j{}^*(\pi, \beta_i, \gamma)]$ is uniquely maximized at the true mean of RPKM, i.e. $\beta_{i\mathbf{0}}$. Suppose the true value of β_i, γ are $\beta_{i\mathbf{0}} = (\beta_{i1\mathbf{0}}, \dots, \beta_{iK\mathbf{0}}), \gamma_{\mathbf{0}} = (\gamma_{1\mathbf{0}, \dots, \gamma_{m\mathbf{0}}})$ and the true value of π is π_0.

A. Identity link function: $\lambda_{ij} = \sum_{k=1}^{K} \beta_{ik} w_{jk} + \sum_{m=1}^{M} \gamma_m v_{jm}$.

The true expectation of Y_{ij} is $\lambda_{ij\mathbf{0}} = \sum_{k=1}^{K} \beta_{ik\mathbf{0}} w_{jk} + \sum_{m=1}^{M} \gamma_{m\mathbf{0}} v_{jm\mathbf{0}}$. Since $E(Y_{ij}|Y_{ij} > 0) = \lambda_{ij}/\pi$ with true value $\lambda_{ij\mathbf{0}}/\pi_0$, it is not hard to show that $E[l_j{}^*(\pi, \beta_i, \gamma)]$ is

$$E[l_j{}^*(\pi, \beta_i, \gamma)] = E[l_j{}^*(\pi, \lambda_{ij})] = (1-\pi_0)\log(1-\pi) + \pi_0 \left(2 \cdot \log(\pi) - \frac{\pi\lambda_{ij\mathbf{0}}}{\pi_0 \lambda_{ij}} - \log(\lambda_{ij})\right) \qquad (7)$$

which is a finite function. By law of large numbers, $L^*(\pi, \beta_i, \gamma)$-the sample mean of $l_j{}^*(\pi, \beta_i, \gamma)$-converges almost surely to $E[l_j{}^*(\pi, \beta_i, \gamma)]$. Next, we need to demonstrate $E[l_j{}^*(\pi, \beta_i, \gamma)]$ being uniquely maximized at $(\pi_0, \beta_{i\mathbf{0}}, \gamma_{\mathbf{0}})$.

We consider the maximizer of $A(\pi, \lambda_{ij}) = E[l_j^*(\pi, \lambda_{ij}, \gamma)] + \pi_0 \log(\lambda_{ij\mathbf{0}})$ instead of $E[l_j^*(\pi, \lambda_{ij})]$.

That is

$$A(\pi, \lambda_{ij}) = E[l_j^*(\pi, \lambda_{ij})] = (1-\pi_0)\log(1-\pi) + \pi_0\,(2\cdot\log(\pi) - \frac{\pi\lambda_{ij\mathbf{0}}}{\pi_0\lambda_{ij}} - \log(\frac{\lambda_{ij}}{\lambda_{ij\mathbf{0}}}))$$

Let $x = \frac{\lambda_{ij}}{\lambda_{ij\mathbf{0}}}$, then

$$A(\pi, x) = A(\pi, \lambda_{ij}) = (1-\pi_0)\log(1-\pi) + 2\pi_0\log(\pi) - (\frac{\pi}{x} + \pi_0\log(x))$$

The first gradient of $A(\pi, x)$ is

$$\frac{\partial A(\pi, x)}{\partial \pi} = -\frac{1-\pi_0}{1-\pi} + \frac{2\pi_0}{\pi} - \frac{1}{x} = 0 \qquad (8)$$

$$\frac{\partial A(\pi, x)}{\partial x} = \frac{\pi}{x^2} - \frac{\pi_0}{x} = 0 \qquad (9)$$

The solution to equations (8) and (9) is $(\pi, x) = (\pi_0, 1)$. The second gradient gives the Hessian matrix

$$H = \begin{bmatrix} \frac{\partial^2 A(\pi,x)}{\partial \pi^2} & \frac{\partial^2 A(\pi,x)}{\partial \pi \partial x} \\ \frac{\partial^2 A(\pi,x)}{\partial \pi \partial x} & \frac{\partial^2 A(\pi,x)}{\partial x^2} \end{bmatrix}_{(\pi_0,1)} = \begin{bmatrix} \frac{1}{1-\pi_0} - \frac{2}{\pi_0} & 1 \\ 1 & -\pi_0 \end{bmatrix}$$

Since $|H| = \frac{1}{1+\pi_0} > 0$, $(\pi_0, 1)$ is the unique maximizer of $A(\pi, x)$. Hence, $(\pi_0, \lambda_{ij\mathbf{0}})$ is the unique solution maximizing $E[l_j^*(\pi, \lambda_{ij})]$.

Lastly, we need to validate that $\lambda_{ij} = \lambda_{ij\mathbf{0}}$ implies $(\beta_i, \gamma) = (\beta_{i\mathbf{0}}, \gamma_{\mathbf{0}})$. According to definitions, the design matrix, (β_i, γ) and λ_{ij} can be written as

$$\begin{bmatrix} w_{11} & \cdots & w_{1K} & v_{11} & \cdots & v_{1M} \\ \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ w_{N1} & \cdots & w_{NK} & v_{N1} & \cdots & v_{NM} \end{bmatrix} \begin{bmatrix} \beta_i \\ \gamma \end{bmatrix} = \begin{bmatrix} \lambda_{i1} \\ \vdots \\ \lambda_{iN} \end{bmatrix} \qquad (10)$$

When $\lambda_{ij} = \lambda_{ij\mathbf{0}}$, equation (10) becomes

$$\begin{bmatrix} w_{11} & \cdots & w_{1K} & v_{11} & \cdots & v_{1M} \\ \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ w_{N1} & \cdots & w_{NK} & v_{N1} & \cdots & v_{NM} \end{bmatrix} \begin{bmatrix} \beta_i \\ \gamma \end{bmatrix} = \begin{bmatrix} \lambda_{i\mathbf{10}} \\ \vdots \\ \lambda_{iN\mathbf{0}} \end{bmatrix} \qquad (11)$$

and $(\beta_{i\mathbf{0}}, \gamma_{\mathbf{0}})$ is a solution to equation (11). Since the design matrix is of full rank, the solution to equation (9) is unique. Therefore, $\lambda_{ij} = \lambda_{ij\mathbf{0}}$ implies $(\beta_i, \gamma) = (\beta_{i\mathbf{0}}, \gamma_{\mathbf{0}})$, and $(\pi_0, \beta_{i\mathbf{0}}, \gamma_{\mathbf{0}})$ is the unique maximizer of $E[l_j^*(\pi, \beta_i, \gamma)]$. The estimator $(\hat{\pi}, \hat{\beta}_i, \hat{\gamma})_{ZI-QML}$ derived from $l_j^*(\pi, \beta_i, \gamma)$ is asymptotically consistent.

B. Logarithmic link function: $\log(\lambda_{ij}) = \sum_{k=1}^{K} \beta_{ik} w_{jk} + \sum_{m=1}^{M} \gamma_m v_{jm}$.

The true expectation of Y_{ij} is $\lambda_{ij\mathbf{0}} = e^{\sum_{k=1}^{K} \beta_{ik\mathbf{0}} w_{jk} + \sum_{m=1}^{M} \gamma_{m\mathbf{0}} v_{jm\mathbf{0}}}$. Similarly, we can derive $E[l_j^*(\pi, \lambda_{ij})]$ as equation (7) and $L^*(\pi, \beta_i)$ converges almost surely to $E[l_j^*(\pi, \lambda_{ij})]$, which is uniquely maximized at $(\pi_0, \lambda_{ij\mathbf{0}})$. Similar to the above proof, $E[l_j^*(\pi, \beta_{ik'}, \gamma)]$ is uniquely maximized at $(\pi_0, \beta_{i\mathbf{0}}, \gamma_{\mathbf{0}})$. Hence, consistency still holds for log link function.

Reference

1. Amemiya, T. and H.U. Press, *Advanced Econometrics*. 1985: Harvard University Press.
2. Gourieroux, C., A. Monfort, and A. Trognon, *Pseudo Maximum Likelihood Methods: Theory.* Econometrica, 1984. **52**(3): p. 681-700.